Generating Positive Psychosis Symptom Keywords from Electronic Health Records*

Natalia Viani[1 [0000−0003−2205−2322]], Rashmi Patel[1,2 [0000−0002−9259−8788]], Robert Stewart[1,2 [0000−0002−4435−6397]], and Sumithra Velupillai[1 [0000−0002−4178−2980]]

[1] IoPPN, King's College London, London, UK
[2] South London and Maudsley NHS Foundation Trust, London, UK
{firstname.lastname}@kcl.ac.uk

Abstract. The development of Natural Language Processing (NLP) solutions for information extraction from electronic health records (EHRs) has grown in recent years, as most clinically relevant information in EHRs is documented only in free text. One of the core tasks for any NLP system is to extract clinically relevant concepts such as symptoms. This information can then be used for more complex problems such as determining symptom onset, which requires temporal information. In the mental health domain, comprehensive vocabularies for specific disorders are scarce, and rarely contain keywords that reflect real-world terminology use. We explore the use of embedding techniques to automatically generate lexical variants of psychosis symptoms into vocabularies, that can be used in complex downstream NLP tasks. We study the impact of the underlying text material on generating useful lexical entries, experimenting with different corpora and with unigram/bigram models. We also propose a method to automatically compute thresholds for choosing the most relevant terms. Our main contribution is a systematic study of unsupervised vocabulary generation using different corpora for an understudied clinical use-case. Resulting lexicons are publicly available.

Keywords: Natural language processing · Electronic health records · Embedding models · Schizophrenia

1 Introduction and Background

Secondary healthcare sources such as electronic health records (EHRs) contain a large proportion of text with clinically relevant information. To analyze this

* RS, RP and SV are part-funded by the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship (MR/S003118/1) and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK. NV and SV have received support by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.

information for clinical research, natural language processing (NLP) techniques are needed. In recent years, NLP systems have been developed to process clinical texts and extract relevant information [1]. An essential step for such systems is the identification of relevant entities, such as medications, symptoms, and time expressions, which can be linked to extract more complex constructs (e.g., treatment and symptom onset). Symptom onset extraction is important in the field of mental health, as a longer duration of untreated symptoms can be associated to worse intervention outcomes [2]. In EHRs related to schizophrenia patients, this information is documented in textual notes in a variety of ways. The first step towards extracting symptom onset is the identification of symptom mentions, which can be achieved using a domain-specific vocabulary. In the mental health domain, however, few standardized vocabularies are available for specific diseases and they rarely contain entries that reflect real-world terminology use.

To develop more comprehensive vocabularies, word embedding techniques can be exploited [3] which rely on neural models to automatically learn word representations (in the form of numeric vectors) from large collections of texts. Given their ability to capture semantic similarity, embedding models have been increasingly used to enhance NLP development, especially for general-domain applications and datasets. In the clinical domain, Ye and Fabbri created embedding models trained on multiple types of EHR clinical notes (e.g., Prescription, Problem List), proposing a method to combine them to enhance term discovery [4]. In the mental health domain, Velupillai et al. compared three approaches for vocabulary generation (dictionary search, linguistic rules, and embedding models) from intensive care unit EHRs [5]. Jackson et al. trained embeddings on mental health EHRs from the Clinical Record Interactive Search (CRIS) database [6], to identify concepts related to mental illness symptomatology.

In this paper, we explore unsupervised embedding models to automatically generate variants of psychosis symptoms indicative of disease onset. Our aim is to generate comprehensive use-case specific lexicons that could be used to solve complex information extraction tasks. Our long-term goal is to identify symptom onset in clinical notes for patients with a diagnosis of schizophrenia. In particular, we study how the choice of the underlying text material impacts the generation of useful terms, comparing four different input corpora and experimenting with bigram models (where frequent word pairs are mapped to a single vector). Moreover, we propose a method to automatically compute appropriate thresholds for choosing the most relevant terms from each model.

2 Materials and Methods

We use data from the CRIS database[3], which gathers anonymized patient information from the EHR system used at the South London and Maudsley NHS Foundation Trust (SLaM) [7].

In our embedding experiments, we trained different (unigram and bigram) models on: 1) Use-case specific EHR texts (*CRIS_specific*, 23.3m words) from

[3] Ethical approval for secondary analysis: Oxford REC C, reference 18/SC/0372.

early psychosis intervention services; 2) Institution-specific discharge summaries (*CRIS_general*, 23.6m words) for all mental health disorders (not restricted to psychosis); 3) External clinical texts from MIMIC II [8] (*MIMIC*, 187.4m words), i.e., an intensive care unit setting. To train embeddings, we used the gensim implementation of Word2Vec, with the CBOW model[4]. We also experimented with 4) Pre-trained embeddings from MEDLINE/PubMed (*PubMed*, 3.6bn words): we used the available models off-the-shelf (only unigram), without re-training [9].

We considered an initial list of keywords from a comprehensive mental health vocabulary [6]. Two psychiatrists reviewed the most frequent vocabulary terms found in *CRIS_specific*, and selected only those that were relevant to identify symptom onset, e.g., *positive* psychosis symptoms. This led to a list of 26 terms - 7 unigrams (e.g., hallucinations), 14 bigrams (e.g., persecutory ideas), and 5 trigrams (e.g., loosening of associations) - which were used for vocabulary generation. For each model, we considered the most similar terms with respect to these keywords (highest vector cosine similarity), and the terms with low Levenshtein distance, i.e., the edit difference (to capture misspellings).

To automatically compute a similarity threshold for each model, we relied on the "elbow" method proposed by Ye and Fabbri [4], which searches for a keyword-specific cutoff point. Given the top **K** similar terms (with decreasing similarity), the method selects the point with maximum distance from the curve connecting the two end-point similarities. In our case, the method was applied to the overall list resulting from all keywords, thus obtaining a model-specific threshold. Since the elbow threshold can change depending on **K**, we automatically computed an optimal value for each model: we tested all **K** values from 50 to 200 and looked at the greatest drop in the resulting elbow threshold.

To evaluate the generated vocabularies, two psychiatrists manually classified terms as: 1) Relevant psychosis symptom term (RT); 2) Potentially relevant term (PT); 3) Not relevant term (NT). As real-world clinical text is likely to contain errors, we also manually assessed the amount of misspelled terms (MSP) per vocabulary. To measure agreement between the raters, we computed the proportion of terms classified with the same label (Ac). Given the nature of this evaluation, other agreement measures (e.g., Cohen's κ) were deemed inappropriate.

3 Results

Fig. 1 shows Venn diagrams for the vocabularies generated by each model. Table 1 reports model-specific results: the number of found original keywords (Keywords), the selected **K**, the vocabulary size, and the number of misspellings (MSP). We report the number of terms classified as relevant by both (RT) or at least one (RT*) rater (the most useful terms for our use-case), and the number of terms classified as PT/NT by both. We also report accuracy on all terms (Ac-all) vs on RT terms only (Ac-RT). Examples of RT were variants (*hallucinatory*), misspellings (*hallicinations*), and specific bigrams (*auditory hallucinations*).

[4] From: https://pypi.org/project/gensim/. Implementation details (preprocessing, parameters) available at: https://github.com/medesto/psychosis-symptom-keywords.

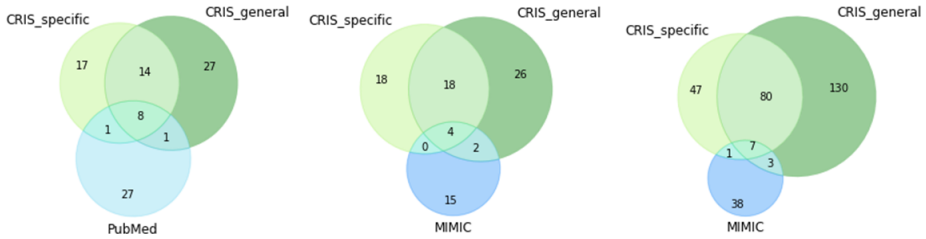

Fig. 1. Venn diagrams for unigram models (center and left) and bigram models (right).

Table 1. Manual evaluation results

Corpus	Model	Keywords	K	All terms	MSP	RT*	RT	PT	NT	Ac-all	Ac-RT
CRIS_specific	unigram	7/7	90	40	0	4	3	12	6	53%	75%
CRIS_general	unigram	7/7	60	50	10	15	13	15	4	64%	87%
MIMIC	unigram	5/7	60	21	1	5	4	2	0	29%	80%
PubMed	unigram	7/7	90	37	4	8	6	11	3	54%	75%
CRIS_specific	bigram	21/21	100	135	0	47	40	44	9	69%	85%
CRIS_general	bigram	19/21	160	220	8	70	58	57	20	61%	83%
MIMIC	bigram	6/21	120	49	1	7	7	6	5	37%	100%

4 Discussion and Conclusion

Generating vocabularies that reflect real-world terminology use is needed to facilitate complex NLP tasks. Moreover, sharing comprehensive lexical resources is an important step to support research in the NLP community. Our main contribution is a systematic study of unsupervised vocabulary generation using different corpora for an understudied clinical use-case. In addition, we proposed a method to automatically compute thresholds to select useful terms from embedding models. All developed resources (vocabularies and evaluations) are made publicly available on our github repository.

A first observation on the impact of corpus selection regards the size of generated vocabularies (Fig. 1). Despite *PubMed* being the largest corpus, the resulting list of terms was comparable to the other models. Also in the case of *MIMIC*, generated vocabularies were relatively small - partially due to some missing original keywords. This observation confirms that larger corpus sizes do not necessarily lead to more useful embedding models [10]. When comparing the two CRIS datasets, the *CRIS_general* model led to a larger vocabulary, especially in the bigram setting (220 vs 135). Interestingly, the proportion of bigram terms was actually higher in the *CRIS_specific* vocabulary (60.7% vs 50.9%). As for misspellings, while the *CRIS_specific* model did not find any entry of interest, the other considered datasets were useful (in particular *CRIS_general*).

As regards the manual evaluation process, the proportion of RT* terms was relatively small, with the most promising results obtained with *CRIS_general* unigram (30%) and *CRIS_specific* bigram (35%). However, most of the remaining terms were classified as potentially useful, which indicates that embedding

models hold potential to capture semantic similarity. It is important to notice that agreement values when considering all labels were lower than those obtained on RT values only. This indicates that it is not straigthforward to distinguish between terms that could be relevant to psychosis and terms that are not relevant at all, which reflects the intrinsic complexity of defining symptoms (and in general the meaning of "relevant") in the mental health domain, hence terminologies. As a starting point, the new RT terms could be successfully reused in support of symptom onset extraction. To improve the proposed methodology/classification, further analysis will be performed on the terms that caused disagreements, with the final aim of developing a psychosis terminology to be linked to SNOMED CT.

As a main limitation of this work, we did not consider different embedding configurations nor n-gram models beyond bigrams. This could have impacted on the small size of RT lists, as single words or word pairs might not be sufficient to identify psychosis symptoms in a definite way (e.g., beliefs, anxiety attacks). More generally, given the intrinsic complexity of this domain, embedding models alone might not be the ideal choice to generate new concepts for specific use-cases. In future work, we will extend our study to take into account more complex models, and we will investigate other ways of modelling the extraction problem.

References

1. Wang, Y., Wang, L., Rastegar-Mojarad, M., et al.: Clinical information extraction applications: A literature review. J Biomed Inf **77**, 34–49 (2018)
2. Kisely, S., Scott, A., Denney, J., Simon, G.: Duration of untreated symptoms in common mental disorders: Association with outcomes. Br J Psychiatry **189**(1), 79–80 (2006)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
4. Ye, C., Fabbri, D.: Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. J Biomed Inf **83**, 63–72 (2018)
5. Velupillai, S., Mowery, D.L., Conway, M., et al.: Vocabulary development to support information extraction of substance abuse from psychiatry notes. In: Proc. BioNLP 2016. pp. 92–101 (2016)
6. Jackson, R., Patel, R., Velupillai, S., et al.: Knowledge discovery for deep phenotyping serious mental illness from electronic mental health records. F1000Research **7** (2018)
7. Perera, G., Broadbent, M., Callard, F., et al.: Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open **6**(3) (2016)
8. Saeed, M., Villarroel, M., Reisner, A.T., et al.: Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical care medicine **39**(5), 952–960 (2011)
9. McDonald, R., Brokos, G.I., Androutsopoulos, I.: Deep relevance ranking using enhanced document-query interactions. In: Proc. EMNLP 2018 (2018)
10. Chiu, B., Crichton, G., Korhonen, A., Pyysalo, S.: How to train good word embeddings for biomedical NLP. In: Proc. BioNLP 2016. pp. 166–174 (2016)